Rev.Phil.Psych. (2018) 9:143–160
DOI 10.1007/s13164-017-0341-8

In Defense of Best-Explanation Debunking Arguments in Moral Philosophy

Jonathon Hricko[1] · **Derek Leben**[2]

Published online: 24 May 2017

Abstract We aim to develop a form of debunking argument according to which an agent's belief is undermined if the reasons she gives in support of her belief are best explained as rationalizations. This approach is a more sophisticated form of what Shaun Nichols has called best-explanation debunking, which he contrasts with process debunking, i.e., debunking by means of showing that a belief has been generated by an epistemically defective process. In order to develop our approach, we identify an example of such a best-explanation debunking argument in Joshua Greene's attack on deontology. After showing that this argument is not an instance of process debunking, we offer our best-explanation approach as a generalization of Greene's argument. Finally, we defend our approach by showing that it is not susceptible to some criticisms that Nichols has leveled against a less sophisticated form of best-explanation debunking.

1 Introduction

Debunking arguments are arguments that purport to demonstrate, not that a belief is false, but that it lacks justification, and is to that extent undermined. For those who think that science matters for ethics, debunking arguments act as a bridge between

✉ Jonathon Hricko
jonathon.hricko@gmail.com

[1] Education Center for Humanities and Social Sciences, National Yang-Ming University, No.155, Linong Street, Sec. 2, Beitou District, Taipei, 112, Taiwan

[2] Department of Philosophy, University of Pittsburgh at Johnstown, 223B Biddle Hall, 450 Schoolhouse Rd., Johnstown, PA, 15904, USA

moral psychology and moral philosophy, since they purport to show that experimental results are relevant when evaluating the justificatory status of moral beliefs and judgments (Joyce 2006; Sinnott-Armstrong 2008; Street 2006). In contrast, those who think that science is irrelevant to ethics often argue that debunking arguments are either unsound or unscientific (Berker 2009; Dean 2010). Our aim in this paper is to develop and defend a particular kind of debunking argument.

The kind of debunking argument we'll defend is a more sophisticated form of what Nichols (2014, pp. 729–730) calls best-explanation debunking. On the best explanation approach that Nichols considers, a belief is undermined if the best explanation of why an agent holds that belief does not involve its truth. The best explanation approach that we'll develop involves the further condition that the best explanation of why she holds her belief is unrelated to the reasons she states in support of her belief. The notion of best explanation enters into our approach in another sense as well, since, in such cases, the agent's reasons are best explained as rationalizations, and her belief is thereby undermined.

We'll proceed as follows. In Section 2, we discuss Nichols's criticisms of the less sophisticated form of best-explanation debunking that he considers. In Section 3, we identify an example of the kind of best-explanation debunking argument we wish to defend, which we find in Greene's (2008, 2014) attack on deontology. Greene draws on an impressive body of experimental results in an attempt to undermine deontological judgments and theories. In short, he argues that we can predict when an agent will make a deontological judgment if we know the agent's level of emotional arousal, and that deontological theories are rationalizations of these emotionally-driven judgments. In Section 4, we generalize from Greene's best-explanation debunking argument, and offer an account of what constitutes a best-explanation debunking argument in general. Finally, we return to Nichols's criticisms, and defend our approach by arguing that it is not susceptible to those criticisms.

2 Nichols on Debunking Arguments

In a recent article on debunking arguments in ethics, Nichols (2014) distinguishes between two forms that such arguments can take, which he calls best-explanation debunking and process debunking, respectively. He goes on to argue that the best-explanation approach ought to be rejected in favor of the process approach. Since we aim to develop and defend a form of best-explanation debunking, we'll begin by considering Nichols's arguments.

One form that a debunking argument might take is what Nichols calls best-explanation debunking, which he characterizes as follows: "if the best explanation of the belief that P doesn't involve the truth of P, then the justificatory status of the belief is undermined" (2014, p. 729).[1] Nichols also emphasizes that, on a more careful statement of this approach, the best explanation of the belief that P would not entail that P is true, *or even likely* (2014, pp. 728–729). Without this clarification,

[1] Nichols finds this kind of debunking argument defended in Joyce (2006) and Tersman (2008).

all beliefs about future actions, like Nichols's belief that he will go to the train station tomorrow morning, would admit of debunking. Importantly, on this approach, the explanation of the belief in question is not supposed to be an explanation of the content of the belief, but an explanation of why a given agent holds that belief. Determining what counts as the best explanation is, of course, a vexed issue.[2] Nichols doesn't elaborate, but we'll have a bit more to say about this issue in Section 4.1, where we develop our own form of best-explanation debunking.

Another form that a debunking argument might take is what Nichols calls process debunking, which he characterizes as follows: "If process Q is an epistemically defective basis for coming to believe that P, then insofar as people believe that P as a result of process Q, their belief that P is unjustified" (2014, p. 731). Nichols has in mind psychological processes in particular (2014, p. 727), and he employs the notion of reliability in order to clarify what makes a psychological process epistemically defective (2014, pp. 745–746). Any evidence that the process in question is unreliable, in the sense that it tends to produce false beliefs, is evidence that the process is epistemically defective. And as Nichols notes, the same process may be epistemically defective in one context, but not in another (2014, p. 733).

Nichols appeals to two kinds of examples in order to argue that process debunking is preferable to best-explanation debunking. The first kind concerns experts who turn out to be wrong about matters in their area of expertise. His examples are Lavoisier, who believed in caloric, a subtle fluid that he hypothesized as the cause of heat; and Frege, who believed in the axiom of comprehension, which leads to Russell's paradox (2014, p. 729). Nichols's basic idea is that Lavoisier and Frege may have been wrong, as indeed they were. But given that they were experts, their beliefs at the time were surely justified. So while one might argue against their beliefs in a number of ways, debunking their beliefs by undermining their justificatory status should not be an option. But, Nichols argues, if one opts for the best-explanation approach, Lavoisier's and Frege's beliefs are ideal candidates for debunking. Their beliefs are false, and so the best explanation of their beliefs won't involve the truth of those beliefs. And on the best-explanation approach, that is enough to debunk their beliefs. On the other hand, debunking is not an option on the process approach. Since Lavoisier and Frege were experts, the processes that led them to their beliefs were presumably not epistemically defective (2014, pp. 730–731). These examples, then, provide one consideration in favor of process-debunking, and against best-explanation debunking.

The second kind of example to which Nichols appeals concerns cases in which an agent holds a true belief for the wrong reasons. His example is John, who believes, truly, that there are electrical impulses in his brain, but his basis for that belief is a delusion that the government has implanted a chip in his head (2014, p. 731). Intuitively, debunking should be an option when it comes to John's belief. But it is not if we opt for the best-explanation approach. This is because the best explanation of John's belief will involve the truth of that belief—if John did not have any electrical impulses in his brain, he would be dead, and wouldn't have any beliefs at all. The process approach, on the other hand, provides a way to debunk John's belief,

[2]See Lipton (2004) for an in-depth treatment of this issue.

since it is the result of deranged paranoia—an epistemically defective process. In that case, we have another consideration in favor of process-debunking, and against best-explanation debunking.

On the best-explanation approach, as Nichols characterizes it, a belief is undermined if the best explanation of that belief does not involve its truth. Nichols's examples, taken collectively, suggest that this condition is neither necessary nor sufficient for debunking. The John case shows that it is unnecessary, and the Lavoisier and Frege cases show that it is insufficient.

While the best explanation approach that Nichols considers is, indeed, deeply flawed, we aim to develop a more sophisticated form of that approach that is not susceptible to the criticisms that Nichols raises in his paper. In Section 4.1, we'll put forward our approach as a sufficient condition for debunking a belief, though not a necessary one. And in Section 4.2, we'll return to the Lavoisier and Frege cases, and argue that our approach fares better than the best-explanation approach that Nichols considers. Since we aren't offering a necessary condition, we're committed to the claim that not all debunking proceeds via the best-explanation approach that we'll advocate. Indeed, our best-explanation approach can be maintained alongside the process approach that Nichols defends. Importantly, though, our best-explanation approach does not ultimately collapse into a kind of process debunking, and so another claim that we'll defend is that the two approaches are, indeed, distinct.

3 Greene's Debunking Argument(s) Against Deontology

In order to develop our best-explanation approach to debunking, we'll now consider what we take to be an example of such a debunking argument, which is found in Greene's (2008, 2014) attack on deontology.[3] According to our reconstruction of his attack, Greene is actually offering two independent, but closely related, debunking arguments. And though he neither distinguishes them nor labels them as such, one is a process debunking argument in Nichols's sense, and the other is a kind of best-explanation debunking argument. In this section, we'll reconstruct both arguments in a fair amount of detail, and we'll do so for two reasons. First of all, while Greene's arguments are well-known, they have not always been well-understood.[4] Secondly, the details matter, since we take one of Greene's arguments to be a paradigm instance of a more sophisticated kind of best-explanation debunking argument, one that ultimately differs from the kind that Nichols criticizes, and from process debunking.

3.1 Psychological Processes and Moral Judgments

Greene is a proponent of the idea that science matters for ethics. As he puts it:

[3]We'll focus on deontology, as Greene usually does. But it's worth noting that Greene intends his arguments to extend to all non-consequentialist moral theories. See Greene (2008, pp. 75–76) and Greene (2008, p. 725).

[4]To take one example, Berker (2009, pp. 315–316) admittedly finds it difficult to determine which arguments Greene is making.

> Science can advance ethics by revealing the hidden inner workings of our moral judgments, especially the ones we make intuitively. Once those inner workings are revealed we may have less confidence in some of our judgments and the ethical theories that are (explicitly or implicitly) based on them. (2014, pp. 695–696)

The first step in both of Greene's debunking arguments involves drawing upon science in order to identify these hidden inner workings, i.e., various psychological and neural processes. And the targets of both debunking arguments are twofold. First of all, Greene aims to undermine certain kinds of moral judgments, specifically, those characteristic of deontology. Secondly, he aims to undermine certain kinds of moral theories, i.e., the deontological ones.[5]

Greene has a particular way of understanding moral theories and the judgments that they purport to justify. He distinguishes between what he calls characteristically deontological judgments and characteristically consequentialist judgments (2008, pp. 38–39; 2014, pp. 699–700). Characteristically deontological judgments are those judgments that are most easily justified by deontological theories (e.g., judgments that the ends don't warrant violating rights or duties). Characteristically consequentialist judgments, on the other hand, are those judgments that are most easily justified by consequentialist theories (e.g., judgments in favor of saving more lives, regardless of what rights and duties may be violated). It's worth noting that some judgments may be neither characteristically deontological nor characteristically consequentialist, i.e., those that are easily justified by both kinds of theory, or by neither kind of theory.

We can illustrate the difference between the two kinds of judgments, as Greene often does, in terms of the trolley cases that he employs in his experimental work (2008, p. 39; 2014, p. 700).[6] In the *switch* case, a runaway trolley will run over, and thereby kill, five people, unless the agent throws a switch, and diverts the trolley so that it will run over and kill one person. Most people judge that it is permissible for the agent to throw the switch. This is a characteristically consequentialist judgment, since it is easily justified in terms of maximizing utility—better to save more lives, if possible. Importantly, a deontologist could make such a judgment and attempt to justify it by appeal to some deontological theory. Greene's point is just that justifying it will be more difficult, and more complicated, than it would be for a consequentialist.

Deontologists have an easier time with the judgment most often made regarding another case—the *footbridge* case. Once again, a runaway trolley is about to run over, and thereby kill, five people. This time, though, the agent has to consider whether or not to push a very large man off of a bridge onto the trolley tracks, thereby killing the man, but stopping the trolley and saving the five in the process. Most people judge that it is impermissible to push the man off of the bridge. This judgment is a characteristically deontological judgment, since it is easily justified in terms of

[5] Strictly speaking, the targets of debunking arguments are beliefs. However, in what follows, we'll talk of debunking judgments and theories. And we'll take it as understood that judgments are a particular kind of belief, and that talk of debunking theories is short for debunking beliefs that are either about theories, or that are generated by theories.

[6] For Greene's experimental work, see, e.g., Greene et al. (2001). For the introduction of the trolley cases in the philosophical literature, see Foot (1967) and Thomson (1976, 1985).

rights and duties—pushing the man is an impermissible violation of his rights. Once again, a consequentialist could make such a judgment, and could attempt to justify it by appealing to some consequentialist theory. But it will be more difficult, and more complicated, than simply appealing to the rights and duties posited by some deontological theory.

Once one distinguishes between such judgments, one can raise an empirical question about the psychological processes that underlie those judgments. There are two such processes that Greene distinguishes in his work: emotional processes, and cognitive processes that involve more controlled, conscious reasoning (2008, pp. 40–41; 2014, pp. 696–698). Emotional processes are automatic, fast, and rather inflexible, and have their source in brain regions like the ventral striatum and ventromedial prefrontal cortex. These processes tend to produce behavior in the here-and-now—e.g., eating that calorie-rich food, and getting away from that tiger. Cognitive processes, on the other hand, are much more flexible, but also much slower. They have their source in brain regions like the dorsolateral prefrontal cortex, and they tend to produce behavior that takes future goals and plans into consideration—e.g., not eating that calorie-rich food now, so that you'll be thinner in the future.

Greene's view of the psychological processes that underlie our moral judgments is encapsulated in what he calls the "Central Tension Principle," which is the most important, and most controversial, claim involved in his dual-process theory of moral judgment. He states the principle as follows:

> Characteristically deontological judgments are preferentially supported by automatic emotional responses, while characteristically consequentialist judgments are preferentially supported by conscious reasoning and allied processes of cognitive control. (2014, p. 699)

Greene discusses a host of evidence for this principle (2008, pp. 41–58; 2014, pp. 700–708). This evidence includes his own experimental work, which involves the aforementioned trolley cases. When confronted with the *footbridge* case, people tend to make the characteristically deontological judgment that it is impermissible to push the man off of the bridge. Greene's experimental results show greater activity in brain regions associated with emotional processing when people make this judgment. On the other hand, when confronted with the *switch* case, people tend to make the characteristically consequentialist judgment that it is permissible to throw the switch. Greene's results also show greater activity in brain regions associated with conscious reasoning when people make this judgment. Here we've just sketched two of Greene's results in a very rough way, and it's worth emphasizing that the evidence in favor of the Central Tension Principle goes well beyond these results. In fact, it goes well beyond the work that Greene and his collaborators have conducted, and includes an impressive body of experimental results from both psychology and neuroscience.

In what follows, we'll assume that the evidence does, indeed, provide strong support for the Central Tension Principle. At this point, we'll distinguish between the two debunking arguments that Greene uses it to support.

3.2 The Best-Explanation Debunking Argument

Greene draws on the emotional origins of our characteristically deontological judgments in order to argue that "the phenomenon of rationalist deontological philosophy is best explained as a rationalization of evolved emotional intuition" (2008, p. 72). This claim forms the basis of his best-explanation debunking argument.

In order to clarify the notion of 'rationalization' that he relies on, Greene (2008, p. 67) considers the case of Alice, who has gone on many dates, after which she reports to you her judgments regarding the people she's dated. Her stated reasons for liking some of these people include such things as kindness, good sense of humor, etc. And her stated reasons for disliking others include such things as arrogance, poor intelligence, etc. You decide to input the data regarding Alice's dates into your statistics software. And you find that Alice has never approved of anyone less than six-foot-four-inches tall, and has never rejected anyone who is six-foot-four-inches tall or taller. While Alice never lists height among her reasons for liking or disliking the people she's dated, you've found that height is a near perfect predictor of whether Alice will approve of those people. Essentially, Alice has a height fetish, and her reasons are rationalizations.

The lesson that Greene draws from the Alice case is that "it's possible to spot a rationalizer without picking apart the rationalizer's reasoning" (2008, p. 67). More specifically, he identifies two things one must do in order to spot a rationalizer:

> First, you have to find a factor that predicts the rationalizer's judgments. Second, you have to show that the factor that predicts the rationalizer's judgments is not plausibly related to the factors that according to the rationalizer are the bases for his or her judgments. (2008, pp. 67–68)

In such cases, an agent's stated reasons for her judgments are best explained as rationalizations, and identifying them as rationalizations undermines those reasons, as well as the judgments that they were meant to support. In short, this argument undermines the rationalizer's judgments by showing that her reasons for making those judgments aren't what she thought they were.

How, then, is Greene's argument supposed to undermine characteristically deontological judgments and the deontological theories that are meant to justify them? To begin with, Greene is concerned with "rationalist versions of deontology such as Kant's; i.e., the ones according to which characteristically deontological moral judgments are justified in terms of abstract theories of rights, duties, etc." (2008, p. 68). The rationalizers, then, are proponents of such theories, i.e., rationalist deontologists. According to the Central Tension Principle, emotion provides the best explanation of how such judgments are actually produced. Emotion is completely unrelated to the reasons that such deontologists give in favor of their judgments. Indeed, in the case of Kant and Kantians, they are explicitly disavowed as good reasons for moral judgments. Hence, their reasons are really just rationalizations, and their characteristically deontological judgments are undermined.

Once again, we can turn to the trolley cases for an illustration.[7] The *footbridge* case elicits a strong emotional response, and people tend to make the characteristically deontological judgment that it's impermissible to push the man off of the bridge. Deontologists justify this judgment by appealing to, say, the fact that the man on the bridge has an inviolable right that you not push him. On the other hand, the *switch* case does not elicit a strong emotional response, and people tend to make the characteristically consequentialist judgment that it's permissible to throw the switch. Deontologists attempt to justify this judgment as well, by appealing to, say, the fact that you violate no one's right by throwing a switch—you're permitted, but not required, to throw it. In the course of attempting to justify both characteristically deontological and characteristically consequentialist judgments, their theories can become quite complicated. But while their theories are complicated, Greene's point is rather simple: It's only when the emotions are sufficiently engaged that deontologists invoke rights and duties to single out particular actions as required or forbidden.

This kind of coincidence calls out for explanation, and as Greene points out, rationalist deontologists have a difficult time explaining why our emotions deliver the correct judgments in such cases (2008, p. 69). Since they're *rationalists*, they have to do so in a way that avoids making those emotions the justificatory grounds of those judgments. While it may not be impossible to deliver such an explanation, the burden of proof is on deontologists to do so. And until they do so, they'll have to leave the relationship between their judgments and the emotions as an unexplained coincidence.

Greene's own explanation of this coincidence is that the theories that deontologists develop are best explained as rationalizations of their emotionally-driven deontological judgments. As a result, their characteristically deontological judgments have been undermined. Moreover, Greene (2014, pp. 718–721) argues that the principles that make up deontological theories are best thought of as concise summaries of such judgments. If Greene is correct that the judgments are thereby justifying the theories, then undermining the judgments has the effect of undermining the theories as well.

However, this move from judgments to theories is not uncontroversial. Kahane and Shackel (2010, pp. 572-580) object to this move on the grounds that Greene's studies do not warrant any conclusions about theories since the relationship between judgments and theories is not at all straightforward. Consider the judgment that it is impermissible to push the man off of the bridge in the *footbridge* case. Greene classifies this judgment as a characteristically deontological judgment. But as Kahane and Shackel observe, a consequentialist might make the same judgment on the grounds that it would maximize utility; and a deontologist might make the opposite judgment on the grounds that the man's right to life is outweighed by other moral considerations. In that case, we cannot use the fact that a person makes what Greene calls characteristically deontological judgments in order to conclude that she is committed to a deontological theory. Because of this disconnect between judgments and

[7]The justifications that we discuss in this paragraph are drawn from Thomson's (1985) proposed solution to the trolley problem. Greene (2008, p. 68) is clear that she counts as a rationalist deontologist for his purposes.

theories, Kahane and Shackel object to Greene's labeling of various judgments as characteristically deontological or characteristically consequentialist. Without the distinction between characteristically deontological and characteristically consequentialist judgments, Greene cannot even state the Central Tension Principle, let alone employ it to undermine deontological theories. This disconnect also casts doubt on Greene's claim that deontological theories are concise summaries of characteristically deontological judgments. After all, such theories tend to have a top-down structure instead of a bottom-up one that begins from judgments about particular cases.

In our view, this objection rests on a misunderstanding of Greene's distinction between characteristically deontological and characteristically consequentialist judgments. Kahane and Shackel claim that "Greene initially suggests that we define '[consequentialist]' and 'deontological' judgements as referring to the judgments typically made by [consequentialist] or deontological philosophers" (2010, p. 578). But as we emphasized in Section 3.1, Greene does not define these judgments in this way. Instead, he distinguishes between these two types of judgments in terms of the kinds of moral theories that would most easily justify them (deontological or consequentialist); and he admits that deontologists sometimes make characteristically consequentialist judgments, and that consequentialists sometimes make characteristically deontological judgments. Moreover, Greene never infers that the participants in his studies who make characteristically deontological judgments are committed to a deontological theory. While Greene himself has commented on Kahane and Shackel's misunderstanding (2014, p. 699), he doesn't discuss its implications for their objection. But once we keep in mind how Greene distinguishes between these two types of judgments, it's clear that a judgment can be connected to a certain kind of moral theory (the kind that most easily justifies it) without pointing unambiguously to some particular theory in the ethics literature. In that case, labeling particular judgments as either characteristically deontological or characteristically consequentialist is no misnomer. The Central Tension Principle is safe from Kahane and Shackel's objection. And although deontologists often present their theories as having a top-down structure, Greene's suggestion that such theories are really crafted to accommodate characteristically deontological judgments still has some bite to it.

It's worth emphasizing that the best-explanation debunking argument that Greene presents differs from both kinds of arguments that Nichols considers. On the best-explanation approach that Nichols considers, the best explanation of the belief debunks that belief, provided that the explanation doesn't involve the truth of the belief. But Greene's argument has a further requirement, namely, that the best explanation of the way in which an agent's judgments are formed is unrelated to the agent's stated reasons. In such cases, the notion of best explanation enters into the argument in a second way, since the agent's stated reasons are best explained as rationalizations. Moreover, the soundness of Greene's argument doesn't require emotional processes to be epistemically defective. In fact, Greene's argument goes through even if we're ignorant about whether emotional processes are reliable when it comes to producing moral judgments. This argument would only count as a process debunking argument, in Nichols's sense, if it involves the claim that emotional processes are epistemically defective.

3.3 The Process Debunking Argument

To be sure, Greene does hold that the emotional processes that underlie our characteristically deontological judgments are epistemically defective, and he develops a related but distinct debunking argument on these grounds. This argument is a process debunking argument in Nichols's sense. We'll now discuss this argument, in order to show that it is distinct from the best-explanation debunking argument discussed above.

Following Leben (2014, pp. 337–339), there are two ways that Greene goes about arguing for the claim that emotional processes are unreliable: the argument from irrelevant factors and the redundancy argument. Greene puts the basic idea of the irrelevant factors argument as follows: "our distinctively deontological moral intuitions . . . reflect the influence of morally irrelevant factors and are therefore unlikely to track the moral truth" (2008, p. 70). This is because characteristically deontological judgments are driven by the emotions, which, in turn, are responsive to whether a harm is "up close and personal" (2008, p. 70). Once again, we'll illustrate the basic idea in terms of the trolley cases. Pushing the man in the *footbridge* case is an example of a personal harm, and such harms tend to elicit a strong emotional response; on the other hand, throwing the switch in the *switch* case is an example of an impersonal harm, and such harms tend not to elicit such an emotional response (2008, p. 43). More generally, personal harms differ from impersonal harms in terms of the spatial distance between the agent and the victim; and in terms of whether the agent harms the victim by means of "personal force" (e.g., by pushing the victim), or by means of something else (e.g., by throwing a switch) (2014, pp. 709–710). Greene argues that it is just obvious, once one reflects on it, that *mere* spatial distance and *mere* personal force are morally irrelevant (2014, p. 713). While some ethicists may hold that these factors are reliable correlates of morally *relevant* factors, Greene holds that even those ethicists must admit that spatial distance and personal force, on their own, are morally irrelevant. Greene's claim, then, is that, because our emotions are responsive to these morally irrelevant factors, they are unreliable when it comes to producing moral judgments.

Greene supplies another reason why deontological judgments are "unlikely to track the moral truth" (2008, p. 70), and this reason forms the basis of the redundancy argument. In short, the reason is that there is an evolutionary explanation of why it is that our emotions are triggered by personal harms, but not by impersonal harms (2008, pp. 43, 70–72).[8] Personal harms have been with us for much longer than impersonal harms. In order to regulate behavior and ensure cooperation, humans evolved a kind of automatic, emotional response to personal harms. But since impersonal harms are somewhat newer, there was no opportunity for an evolved response to such harms. Importantly, then, our emotional response to personal harms was selected

[8]Similar evolutionary debunking arguments have been put forward by Joyce (2001, 2006), Rosenberg (2011), Ruse (1986), and Street (2006), among others. Most of these authors claim that discovering the evolutionary history of the processes that produce moral judgments undermines *all* moral judgments, or at least those that presuppose some mind-independent moral truth. In contrast, Greene targets only characteristically deontological judgments.

for because of its role in functions that had nothing to do with detecting moral truths. Leben (2014, pp. 337–338) labels this argument a redundancy argument because the evolutionary explanation shows that emotional processes would generate the same responses whether or not the deontological judgments they produce were true. The truth of such judgments, as well as the existence of the deontologist's rights and duties, are therefore redundant, in which case our emotional processes fail to track the moral truth. Thus, Greene concludes that those processes are unreliable when it comes to producing moral judgments.

To sum up: Greene's process debunking argument amounts to his defense of the claim that the factor to which the best explanation appeals (i.e., emotion) is an epistemically defective psychological process. And his defense of this latter claim comes from the argument from irrelevant factors, and from the redundancy argument. The process debunking argument can be understood as consisting of the best-explanation debunking argument, supplemented with the claim that emotional processes are epistemically defective when it comes to producing moral judgments. The process debunking argument and the best-explanation debunking argument are therefore closely related, but logically distinct.

3.4 A Taxonomy of Debunking Arguments

Given the close relationship between the process debunking argument and the best-explanation debunking argument, one might wonder why we distinguish between them. In order to explain why, we propose a broad taxonomy of the debunking arguments described thus far, illustrated in Fig. 1.

It is useful to distinguish between the process debunking argument and the best-explanation debunking argument because, in our view, there are ultimately two different methods by which Greene undermines belief in deontological judgments and theories. The first method is to show that deontological theories are rationalizations of emotionally-driven deontological judgments. This method is encapsulated in the best-explanation debunking argument. The second method is to show that the emotional processes that underlie characteristically deontological judgments are epistemically defective. This method is encapsulated in the process debunking argument.

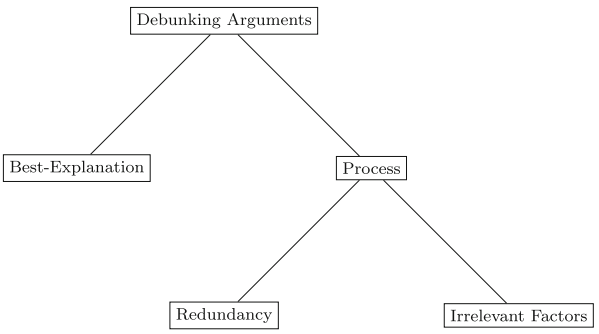

Fig. 1 A taxonomy of debunking arguments

One might wonder whether one method renders the other method redundant. In our view, both arguments are doing important work, and this is because undermining beliefs is a matter of degree. Greene's best-explanation debunking argument undermines deontological judgments and theories to a degree, and his process debunking argument undermines such judgments and theories to a further degree.

In Section 3.3, we introduced the redundancy argument and the argument from irrelevant factors as presenting two ways in which emotional processes may be epistemically defective. Given that characterization of these two arguments, it's clear that each is a distinct kind of process debunking argument, represented as such in the taxonomy in Fig. 1. In order to distinguish these arguments, it's useful to conceive of reliability in terms of whether a process yields false positives and/or false negatives, or alternatively, whether it commits something analogous to the type I and II errors familiar from the social sciences. As Leben (2014, p. 338) makes clear, the redundancy argument shows that emotional processes are unreliable in the sense that they are likely to yield false positives. We would expect emotionally-driven judgments that there are such-and-such rights and duties regardless of whether any such rights and duties exist. In contrast, the argument from irrelevant factors shows that emotional processes are unreliable by showing that they are responsive to the morally irrelevant factor of whether a harm is personal or not. Once again, we would expect false positives regarding the existence of certain rights and duties. And those who think that rights and duties do exist might even expect false negatives regarding the non-existence of certain rights and duties.

So far, we've employed the taxonomy in Fig. 1 in order to distinguish the debunking arguments that Greene employs in his attack on deontology. It should be clear that we take Greene's arguments to be particular instances of the more general kinds of debunking arguments that we've labeled in the figure. In the remainder of the paper, we'll focus on one of those more general kinds, namely, best-explanation debunking arguments.

4 In Defense of Best-Explanation Debunking

The upshot of the preceding discussion is that Greene's best-explanation debunking argument is distinct from the form of best-explanation debunking that Nichols considers, and from process debunking. Our final task is to generalize from the case of Greene's argument in order to develop and defend our approach to best-explanation debunking.

4.1 Generalizing From Greene's Argument

Our approach to best-explanation debunking results from combining a generalized form of Greene's best-explanation debunking argument with the best-explanation approach that Nichols considers. Greene's approach debunks beliefs by showing that an agent's stated reasons for holding those beliefs are best explained as rationalizations. On the approach that Nichols considers, if the best explanation of why an agent holds a belief does not involve the truth of that belief, the agent's belief is debunked.

Our own approach is encapsulated in the following three conditions, which suffice for an explanation E to debunk an agent A's belief(s) B:

1. E best explains why A holds B in terms of A's real reasons RR for holding B.
2. Neither E nor RR entails that B is true or likely.
3. A's stated reasons SR for holding B are unrelated to E, i.e., $SR \neq RR$.

When these three conditions are satisfied, E debunks B, and it does so even if E appeals to a psychological process that is not known to be epistemically defective. Importantly, we take these three conditions to be jointly sufficient, though not necessary, for debunking a belief. They are the conditions under which SR is best explained as a rationalization, thereby debunking B. The notion of best explanation plays a dual role here, in the sense that we need the *best explanation* of why A holds B, and in the sense that SR is *best explained* as a rationalization. The intuitive idea that we aim to capture with these conditions is that a rationalizer's reasons for holding a belief are not what she thinks they are, and pointing this out is sufficient to debunk her beliefs.

We'll now go through each condition individually in more detail. The first condition, which is shared with the best-explanation approach that Nichols considers, appeals to E being the best explanation of why A holds B. What makes an explanation the best is, to a large extent, relative to the branch of science in which it is put forward, and to the phenomena it attempts to explain. Since we're concerned with explaining the production of beliefs, the relevant branches of science are the cognitive and behavioral sciences. Following Greene (2008, pp. 67–68), we take it that two features that make such an explanation better than its competitors are prediction and simplicity: Can we use statistical techniques to identify a single factor that predicts an agent's beliefs? If so, does that factor yield an explanation that is simpler than the agent's stated reasons? Given that explanation in the cognitive and behavioral sciences is a contentious issue,[9] we don't want to claim that these two features are sufficient for concluding that one explanation is better than another. But regardless of how one understands explanation in the cognitive and behavioral sciences, it will have to be the case that prediction and simplicity play a significant role in determining which explanation is the best.

The second condition specifies that E does not entail that B is true or likely. Like the first condition, this condition is shared with the best-explanation approach that Nichols considers. Since we understand this condition in more or less the same way that Nichols does, no further comment is needed.

The third condition requires that A's stated reasons SR for holding B are unrelated to E, i.e., that $SR \neq RR$. It is this condition that captures Greene's insight regarding rationalization, and it thereby distinguishes our approach most clearly from both of the approaches that Nichols considers. Greene's cases of the imaginary Alice and the historical Kant suggest one rather straightforward way of determining whether SR and E are related. Alice does not recognize height as one of the reasons that she rejects her suitors, and Kant explicitly denies that emotion can justify a moral judgment. More generally, in order to determine whether SR and E are related, it

[9] See Gervais (2015) for an up-to-date account of the debate.

often suffices to determine whether A denies (or would deny, if asked) that SR and E are related. While this may not work in all cases, Greene's examples show that it does work sometimes, and this is enough for our purposes.

We claim that, if these three conditions are satisfied, then E debunks A's belief(s) B. This is the case because there is something epistemically bad about beliefs originating from a source different from the one an agent explicitly cites. Thus, the justificatory status of a belief is lowered as a result of this discovery. As we suggested in Section 3.4, this lowering is a matter of degree. A best-explanation debunking argument may lower the justificatory status of a belief while leaving some degree of justification intact.[10]

It's worth going into more detail regarding what, exactly, is epistemically bad about rationalization. Our claim is that there's a problem inherent in the lack of connection between E and SR. We'll now consider three arguments for this claim.

The first argument is that rationalization is epistemically bad because it prevents an agent from having access to her own reasons. This argument is based on epistemic internalism, according to which an agent is justified in her belief only if she has access to the reasons for holding that belief (BonJour 1980; Conee and Feldman 2001; Steup 1999). Internalists emphasize how a justification is more than just a mechanism that reliably indicates the truth. Rather, agents must have some kind of 'second-order' knowledge about the sources of their beliefs. An internalist would say that discovering that A's stated reasons SR for her belief B are unrelated to her real reasons RR demonstrates that the agent has no account of the connection between RR and B, thus undermining the justification for B. However, epistemic internalism is controversial, and many authors deny that introspective access is required for justification (Goldman 1999; Kornblith 1999; Sosa 2003). Thus, we'll consider two other arguments that support our claim.

The second argument is that rationalization is epistemically bad because it does not open a belief to rational evaluation and criticism. This argument is due to Schwitzgebel and Ellis (2017), who are concerned with a kind of rationalization that differs from the kind that we have in mind. By 'rationalization,' they have in mind cases in which A's real reasons RR have their source in an epistemically defective process (2017, p. 184), which would be sufficient to launch a process debunking argument. We have in mind a more minimal notion of rationalization, which merely requires there to be a mismatch between SR and RR, and which doesn't require that RR has its source in an epistemically defective process. Regarding rationalization in their sense, Schwitzgebel and Ellis (2017, p. 183) describe how being mistaken about the real reasons for one's beliefs can block an essential part of the discursive process of evaluating beliefs:

> There's a type of dialectical critique that is, we think, epistemically important in moral and philosophical reasoning—we might call it *engaged* or *open*

[10]There are a number of ways to make this idea more precise. For example, one might claim that the justificatory status of a belief is lowered in proportion to the extent to which E is a better explanation of why A holds B than SR is, and to the extent to which E and SR are unrelated.

> dialogue—in which one aims to offer to an interlocutor, for the interlocutor's examination and criticism, one's *real reasons* for believing some conclusion. . . . Rationalization disrupts this type of peer critique. One's real basis remains hidden; it's not really up for peer examination, not really exposed to the risk of refutation or repudiation.

Their argument applies equally well to rationalization in our sense. Rationalization closes off the rational structure of a belief, and thereby closes it off from the normal procedures of peer-critique and even self-critique. The point of this argument is that any beliefs that are closed off in this way are undermined by virtue of not participating in the standard practices of justification. There are similarities between this argument and epistemic internalism, but Schwitzgebel and Ellis's argument has a more pragmatic angle. It states that being open to public scrutiny (rather than introspective access) is an essential part of the practice of justifying beliefs.

A third, related, argument is that rationalization is epistemically bad because we rightly do not take a rationalizer's stated reasons very seriously. In the course of presenting their 'peer critique' argument, Schwitzgebel and Ellis (2017, p. 183) claim that when one states one's reasons for some belief,

> One says, "here's why I think P," with the aim of offering considerations in favor of P that simultaneously play two roles: (i) they epistemically support P (at least prima facie); and (ii) acceptance of them is actually causally effective in sustaining one's belief that P is the case.

Spotting a rationalizer is sufficient to cast doubt on (ii) since A's stated reasons SR are not, in fact, the cause of B. As a result, A would believe B regardless of whether SR provides any support for B. Hence, we'd be justified in not taking SR seriously. It may turn out to be the case that SR does, in fact, support B. But until that is shown to be the case, the justificatory status of B is lowered.

It's worth emphasizing our point that, if our three conditions are satisfied, then E debunks B even if E appeals to a psychological process that *is not known* to be epistemically defective. This point is important because there are cases in which E appeals to a psychological process that is not known to be epistemically defective, and so there are cases of debunking that are distinct from process debunking. In such cases, showing that an agent is rationalizing suffices for debunking her beliefs—no appeal to an epistemically defective process is necessary.

One might wonder whether our best-explanation approach is really a form of the process approach that Nichols defends, on the grounds that rationalization is itself an epistemically defective psychological process. To be sure, we do hold that there is a sense in which rationalization is an epistemically defective process. After all, our main point is that showing that an agent is a rationalizer is sufficient to debunk her beliefs. But there is a significant difference between rationalization and the epistemically defective processes with which Nichols is concerned. Nichols is concerned with processes that generate beliefs (2014, p. 730). But rationalization is a process that takes place after beliefs have been generated. Hence, our best-explanation approach ultimately differs from Nichols's process approach.

As opposed to the necessary-and-sufficient conditions for debunking that Nichols considers, our three conditions are merely jointly sufficient. Hence, this approach does not pretend to capture the *only* way debunking works. We admit that process debunking, which we also understand as a sufficient condition for debunking, can be maintained alongside our form of best-explanation debunking. And process debunking may undermine belief to a greater degree than our form of best-explanation debunking can, at least in some cases. That said, it is important to distinguish these two kinds of arguments, and to show that not all debunking proceeds via the identification of an epistemically defective process.

4.2 Defending Against Nichols's Criticisms

We'll now defend our form of best-explanation debunking by showing that it is not susceptible to the criticisms that Nichols raises against the less sophisticated form of best-explanation debunking that he considers. Nichols uses the Lavoisier and Frege cases in order to show that the best-explanation approach that he considers can't be a sufficient condition for debunking. Since we've claimed to present a sufficient condition for debunking, we'll now consider what our account entails regarding these cases. We'll focus on the Lavoisier case in particular, with the understanding that what we have to say about this case applies equally well to the Frege case.

To begin with, we'll briefly recapitulate Nichols's argument. Nichols (2014, pp. 729–730) argues that, since Lavoisier was an expert chemist, his belief in caloric, the subtle fluid that he hypothesized as the cause of heat, was justified. In that case, it shouldn't be possible to debunk his belief by undermining its justificatory status. But the best explanation of why Lavoisier held that belief will not involve the truth of that belief, since that belief is false. And so, on the best-explanation approach that Nichols considers, we're left with the undesirable result that we can, in fact, debunk the beliefs of experts like Lavoisier. In that case, the best-explanation approach that Nichols considers can't be a sufficient condition for debunking.

On the best-explanation approach that we've just presented, it's possible to debunk experts' beliefs, but it takes a lot of work. In short, this is because it's difficult to satisfy the third condition (that the agent's stated reasons for holding a belief are unrelated to the best explanation of why she holds that belief). We take this result to be the desirable one, and the biggest flaw regarding the best-explanation approach that Nichols considers is that it's much too easy to debunk experts' beliefs.

Our best-explanation approach yields the following picture of the Lavoisier case. Lavoisier's stated reasons in favor of his belief in caloric include: his observation of changes of state of various substances (for example, of ice into water), his belief that such changes involved the gradual separation of the particles that make up those substances, and his belief that some additional material substance is the most likely cause of that separation (1965/1789, pp. 1–6). In Lavoisier's case, our best explanation of why he believed in caloric is going to include his stated reasons. And if it does, then those stated reasons will thereby be related to the best explanation of his belief, thus violating the third condition. In that case, we can't debunk Lavoisier's belief.

At this point, one might object that we should be able to tell the same story about rationalist deontologists that we've just told about Lavoisier. After all, just as Lavoisier was an expert chemist, deontologists are expert ethicists. In that case, their stated reasons should surely be included in the best explanation of their characteristically deontological judgments. And so, just as in Lavoisier's case, the third condition is violated.

However, there is a reason that we can't tell the same story about rationalist deontologists that we've told about Lavoisier. When it comes to the former, Greene has identified a better explanation of their beliefs, and that explanation is unrelated to their stated reasons. To be sure, it is, strictly speaking, possible that we'll eventually arrive at a better explanation of Lavoisier's belief in caloric, one that is unrelated to his stated reasons. But it would take some work to show that to be the case, and until one does, we have no reason to think that, in Lavoisier's case, the third condition is satisfied. Hence, we can't tell the same story about rationalist deontologists that we've told about Lavoisier. This is because, in the former case, we have Greene's explanation, and we don't have anything comparable in the latter case.

The important point is that our best-explanation approach is not susceptible to Nichols's criticisms. This is because it makes debunking beliefs a non-trivial task, whereas the best-explanation approach that Nichols considers entails that any false belief is trivially susceptible to debunking. On our view, debunking beliefs is a non-trivial task because we assume that an agent's stated reasons are related to the best explanation of her belief unless we have some compelling evidence to think otherwise. In other words, we assume that the third condition is violated unless we have some positive reason to think that it has been satisfied. Greene's work shows that it is possible, but difficult, to satisfy the third condition. And so, while we can't conclude *a priori* that anyone's belief (including experts like Lavoisier) is entirely immune from debunking, it does require some substantial work to debunk an agent's belief.

5 Conclusion

In this paper, we've developed and defended a form of best-explanation debunking that undermines an agent's belief by showing that agent to be a rationalizer. More specifically, if the agent's stated reasons for holding a given belief are unrelated to the best explanation of why she holds that belief, then her stated reasons are best explained as rationalizations, and her belief is thereby undermined. This form of debunking can co-exist alongside process debunking arguments that debunk by means of identifying an epistemically defective psychological process. But our approach is distinct from process debunking, and we take it to be an important conclusion that one can undermine a belief without showing it to be the result of an epistemically defective process. Our central example, namely, Greene's best-explanation debunking argument, concerns beliefs in the moral domain. But our discussion of Nichols's nonmoral cases suggests that our approach might extend to beliefs that fall within nonmoral domains as well.

Acknowledgements We presented an earlier draft of this material at the Institute of European and American Studies at Academia Sinica. Thanks to the audience there, and especially to Tzu-wei Hung, Ellie Hua Wang, Terence Hua Tai, and Norman Y. Teng for helpful questions and comments. Thanks also to two anonymous reviewers for their helpful suggestions. Much of the work for this paper was completed during Hricko's time as a postdoctoral fellow in the Institute of European and American Studies at Academia Sinica, and so he would like to thank the Institute, and especially his sponsor Jih-Ching Ho.

References

Berker, S. 2009. The normative insignificance of neuroscience. *Philosophy & Public Affairs* 37(4): 293–329.

BonJour, L. 1980. Externalist theories of empirical knowledge. *Midwest Studies in Philosophy* 5: 53–73.

Conee, E., and R. Feldman. 2001. Internalism defended. *American Philosophical Quarterly* 38(1): 1–18.

Dean, R. 2010. Does neuroscience undermine deontological theory? *Neuroethics* 3(1): 43–60.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5: 5–15.

Gervais, R. 2015. Mechanistic and non-mechanistic varieties of dynamical models in cognitive science: Explanatory power, understanding, and the 'mere description' worry. *Synthese* 192(1): 43–66.

Goldman, A. 1999. Internalism exposed. *Journal of Philosophy* 96(6): 271–293.

Greene, J.D. 2008. The secret joke of Kant's soul. In *Moral psychology, vol. 3: The neuroscience of morality: Emotion, disease, and development*, ed. W. Sinnott-Armstrong, 35–79. Cambridge: MIT Press.

Greene, J.D. 2014. Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics* 124(4): 695–726.

Greene, J.D., R.B. Sommerville, L.E. Nystrom, J.M. Darley, and J.D Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537): 2105–2108.

Joyce, R. 2001. *The myth of morality*. Cambridge: Cambridge University Press.

Joyce, R. 2006. *The evolution of morality*. Cambridge: MIT Press.

Kahane, G., and N. Shackel. 2010. Methodological issues in the neuroscience of moral judgement. *Mind & Language* 25(5): 561–582.

Kornblith, H. 1999. Distrusting reason. *Midwest Studies in Philosophy* 23: 181–196.

Lavoisier, A.L. 1965/1789. *Elements of chemistry*. New York: Dover.

Leben, D. 2014. When psychology undermines belief. *Philosophical Psychology* 27(3): 328–350.

Lipton, P. 2004. *Inference to the best explanation*, 2nd edn. London: Routledge.

Nichols, S. 2014. Process debunking and ethics. *Ethics* 124(4): 727–749.

Rosenberg, A. 2011. *The atheist's guide to reality: Enjoying life without illusions*. New York: W. W. Norton & Company.

Ruse, M. 1986. *Taking Darwin seriously: A naturalistic approach to philosophy*. New York: Oxford University Press.

Schwitzgebel, E., and J. Ellis. 2017. Rationalization in moral and philosophical thought. In *Moral Inferences*, eds. J.-F. Bonnefon, and B. Trémolière, 170–190. New York: Routledge Psychology Press.

Sinnott-Armstrong, W. 2008. Framing moral intuitions. In *Moral psychology, vol. 2: The cognitive Science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong, 47–76. Cambridge: MIT Press.

Steup, M. 1999. A defense of internalism. In *The theory of knowledge: Classical and contemporary readings*, 2nd edn, ed. L. Pojman, 373–384. Belmont: Wadsworth Publishing.

Sosa, E. 2003. Beyond internal foundations to external virtues. In *Epistemic justification: Internalism vs. externalism, foundations vs. virtues*, eds. L. BonJour, and E. Sosa, 97–170. Malden: Blackwell.

Street, S. 2006. A Darwinian dilemma for realist theories of value. *Philosophical Studies* 127(1): 109–166.

Tersman, F. 2008. The reliability of moral intuitions: a challenge from neuroscience. *Australasian Journal of Philosophy* 86(3): 389–405.

Thomson, J.J. 1976. Killing, letting die, and the trolley problem. *The Monist* 59(2): 204–217.

Thomson, J.J. 1985. The trolley problem. *The Yale Law Journal* 94(6): 1395–1415.